日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００１年　６月１８日

出願番号
Application Number: 特願２００１－１８３８５６

[ST. 10/C]: [ＪＰ２００１－１８３８５６]

出願人
Applicant(s): 科学技術振興事業団

２００３年１０月２２日

特許庁長官
Commissioner,
Japan Patent Office 今井康夫

出証番号　出証特２００３－３０８７１３６

【書類名】　特許願
【整理番号】　A000103431
【特記事項】　特許法第３０条第１項の規定の適用を受けようとする特許出願
【提出日】　平成13年 6月18日
【あて先】　特許庁長官　殿
【国際特許分類】　G06F 17/00
【発明の名称】　相同性解析システム、相同性解析方法及び相同性解析プログラム
【請求項の数】　11
【発明者】
【住所又は居所】　埼玉県熊谷市久下２０３８－３
【氏名】　篠沢　隆雄
【発明者】
【住所又は居所】　群馬県桐生市西久方町１－５－２３　パークハイツ３０６
【氏名】　堀池　徳祐
【特許出願人】
【識別番号】　396020800
【氏名又は名称】　科学技術振興事業団
【代理人】
【識別番号】　100058479
【弁理士】
【氏名又は名称】　鈴江　武彦
【電話番号】　03-3502-3181
【選任した代理人】
【識別番号】　100092196
【弁理士】
【氏名又は名称】　橋本　良郎

【選任した代理人】

【識別番号】　100091351

【弁理士】

【氏名又は名称】　河野　哲

【選任した代理人】

【識別番号】　100088683

【弁理士】

【氏名又は名称】　中村　誠

【手数料の表示】

【予納台帳番号】　011567

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　相同性解析システム、相同性解析方法及び相同性解析プログラム

【特許請求の範囲】

【請求項１】　解析対象となる解析対象データ群が、前記解析対象データ群とは異なる第１のデータ群あるいは第２のデータ群のいずれに類似するかを解析するための相同性解析システムであって、

前記解析対象データ群と第１のデータ群のそれぞれのデータ群に含まれるデータの相同性を示す第１の相同値ｘを算出するものであって、前記しきい値Ｅをｎ個設定して各しきい値E_i（ｉ＝１，２，…ｎ）毎に第１の相同値x_iを算出する第１の相同値算出手段と、

前記解析対象データ群と第２のデータ群のそれぞれのデータ群に含まれるデータの相同性を示す第２の相同値ｙを算出するものであって、前記しきい値Ｅをｎ個設定して各しきい値E_i（ｉ＝１，２，…ｎ）毎に第２の相同値y_iを算出する第２の相同値算出手段と、

第１の相同値x_i及び第２の相同値y_i並びにしきい値の数ｎとの関係に基づいて、前記解析対象データ群が第１のデータ群又は第２のデータ群のいずれに類似するかを判定する相同性決定手段と

を具備してなることを特徴とする相同性解析システム。

【請求項２】　前記第１の相同値算出手段は、

前記解析対象データ群と第１のデータ群のそれぞれのデータ群に含まれるデータの相同性が所定のしきい値Ｅ以上である場合には相同性有りと判定し、相同性有りとなるデータの数を第１の相同値ｘとして算出し、

前記第２の相同値算出手段は、

前記解析対象データ群と第２のデータ群のそれぞれのデータ群に含まれるデータの相同性が所定のしきい値Ｅ以上である場合には相同性有りと判定し、相同性有りとなるデータの数を第１の相同値ｙとして算出する

ことを特徴とする請求項１に記載の相同性解析システム。

【請求項３】　前記第１のデータ群はn_A個のデータを有し、前記第２のデ

ータ群はｎB個のデータを有する場合、

前記第１の相同値算出手段は、一つのしきい値Ｅｉに対して前記第１のデータ群の各データ毎に第１の相同値ｘｉｊ（ｊ＝１，２，…ｎA）を算出し、その算出されたしきい値Ｅｉについての第１の相同値ｘｉの平均値ｘｉ＿を算出し、

前記第２の相同値算出手段は、一つのしきい値Ｅｉに対して前記第２のデータ群の各データ毎に第１の相同値ｙｉｊ（ｊ＝１，２，…ｎB）を算出し、その算出されたしきい値Ｅｉについての第２の相同値ｙｉの平均値ｙｉ＿を算出し、

前記相同性決定手段は、

【数１】

$$u_i = \sqrt{\frac{1}{n_A + n_B - 2}\left\{\sum_{j=1}^{n_A}(x_{ij} - \overline{x_i})^2 + \sum_{k=1}^{n_B}(y_{ik} - \overline{y_i})^2\right\}}$$

とした場合に、以下の等式

【数２】

$$Z_i^{(1)} = \frac{\overline{x_i} - \overline{y_i}}{u_i} \cdot \sqrt{\frac{n_A \cdot n_B}{n_A + n_B}} \qquad (i = 1,2,\ldots,n)$$

により、第１のデータ群あるいは第２のデータ群のいずれに類似するかを示す相同性決定値Ｚｉ（１）を算出することを特徴とする請求項１に記載の相同性解析システム。

【請求項４】　前記相同性決定値Ｚｉ（１）がｔ分布に従い、自由度αとすると、ｔα（０，１０）よりも大きい場合には前記解析対象データ群は前記第１のデータ群と相同性を持つデータ数が多いと判定する判定結果導出手段をさらに備えることを特徴とする請求項３に記載の相同性解析システム。

【請求項５】　前記相同性決定値Ｚｉ（１）がｔ分布に従い、自由度αとすると、－ｔα（０，１０）よりも小さい場合には前記解析対象データ群は前記第２のデータ群と相同性を持つデータ数が多いと判定する判定結果導出手段をさらに備えることを特徴とする請求項２に記載の相同性解析システム。

【請求項６】　前記判定結果導出手段はさらに、ｓをＺｉ（１）の標準偏差、Ｚｉ（１）＿をＺｉ（１）の平均値とすると、

【数３】

$$Z^{(2)} = \frac{|\overline{Z^{(1)}}| - t_{n_A+n_B-2}(0.10)}{s/\sqrt{n-1}}$$

の等式で示される相同性可否決定値Ｚ（２）を算出し、該相同性可否決定値Ｚ（２）が所定の値t_{n-1}（０，１０）未満の場合には相同性決定値Z_i（１）が有効な値でないと判定することを特徴とする相同性可否判定手段を備えることを特徴とする請求項４又は５に記載の相同性解析システム。

【請求項７】　前記自由度αは、n_A+n_B-2であることを特徴とする請求項４乃至６に記載の相同性解析システム。

【請求項８】　前記第１及び第２の相同値算出手段は、ＢＬＡＳＴ法により相同値x_i及びy_iを算出することを特徴とする請求項１乃至７に記載の相同性解析システム。

【請求項９】　前記解析対象データ群、第１のデータ群及び第２のデータ群は、遺伝子の配列を示すデータであることを特徴とする請求項１乃至８に記載の相同性解析システム。

【請求項１０】　解析対象となる解析対象データ群が、前記解析対象データ群とは異なる第１のデータ群あるいは第２のデータ群のいずれに類似するかを解析するための相同性解析方法であって、

前記解析対象データ群と第１のデータ群のそれぞれのデータ群に含まれるデータの相同性を示す第１の相同値ｘを算出するステップであって、この第１の相同値ｘを、前記しきい値Ｅをｎ個設定して各しきい値E_i（ｉ＝１，２，…ｎ）毎に第１の相同値x_iとして算出するステップと、

前記解析対象データ群と第２のデータ群のそれぞれのデータ群に含まれるデータの相同性を示す第２の相同値ｙを算出するステップであって、この第２の相同値ｙを、前記しきい値Ｅをｎ個設定して各しきい値E_i（ｉ＝１，２，…ｎ）毎に第２の相同値y_iとして算出するステップと、

第１の相同値x_i及び第２の相同値y_i並びにしきい値の数ｎとの関係に基づいて、前記解析対象データ群が第１のデータ群又は第２のデータ群のいずれに類似

するかを判定するステップと

を有することを特徴とする相同性解析システム。

【請求項１１】　解析対象となる解析対象データ群が、前記解析対象データ群とは異なる第１のデータ群あるいは第２のデータ群のいずれに類似するかを解析するための相同性解析プログラムであって、コンピュータを、

前記解析対象データ群と第１のデータ群のそれぞれのデータ群に含まれるデータの相同性を示す第１の相同値ｘを算出するものであって、前記しきい値Ｅをｎ個設定して各しきい値E_i（ｉ＝１，２，…ｎ）毎に第１の相同値x_iを算出する第１の相同値算出手段と、

前記解析対象データ群と第２のデータ群のそれぞれのデータ群に含まれるデータの相同性を示す第２の相同値ｙを算出するものであって、前記しきい値Ｅをｎ個設定して各しきい値E_i（ｉ＝１，２，…ｎ）毎に第２の相同値y_iを算出する第２の相同値算出手段と、

第１の相同値x_i及び第２の相同値y_i並びにしきい値の数ｎとの関係に基づいて、前記解析対象データ群が第１のデータ群又は第２のデータ群のいずれに類似するかを判定する相同性決定手段

として機能させることを特徴とする相同性解析プログラム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、複数のデータからなるデータ群同士の相同性を解析するための相同性解析システム、相同性解析方法及び相同性解析プログラムに関する。

【０００２】

【従来の技術】

遺伝子の類似性を比較する従来の方法として最も普及している方法は、「系統樹解析」である。この方法は、遺伝子の類似性を含めた相互関係を最も厳密に解析することができる。しかし、多数の遺伝子に関する解析を行う場合は、それに要する時間が長い点において大きな欠点を要する。また、生物の固体間の全体的な類似性、すなわち多数の遺伝子のマスとしての相関関係などの解析には、採用

した遺伝子が本当にその生物間の関係を代表しているものであるかという点で、適当な遺伝子を選んでいるとは限らない。遺伝子の相同性・類似性を比較する方法としては、上記の系統樹解析に加えて、ＢＬＡＳＴ法やＦＡＳＴＡ法が代表的である。ここで、ＢＬＡＳＴ法とは、遺伝子のデータバンクの中からある特定の遺伝子と類似性を有する遺伝子を選別するプログラムを用いて解析を行う方法である。

【０００３】

【発明が解決しようとする課題】

上記したようなＢＬＡＳＴ法やＦＡＳＴＡ法では、遺伝子間の類似性の度合い、すなわちどれほど類似性が高いかに応じて、どのような遺伝子がどの程度の類似性があるかということを解析し、遺伝子を選別・特定することができる。しかし、これらの方法は類似性の度合いを設定する領域、すなわちしきい値を固定しているために、他の類似性の度合いにおいて、どのような遺伝子が存在するかなどの解析を行う点で不十分であった。

【０００４】

本発明は上記課題を解決するためになされたもので、その目的とするところは、多数の遺伝子同士を定量的にかつ正確に比較することが可能な相同性解析システム、相同性解析方法及び相同性解析プログラムを提供することにある。

【０００５】

【課題を解決するための手段】

この発明のある観点によれば、解析対象となる解析対象データ群が、前記解析対象データ群とは異なる第１のデータ群あるいは第２のデータ群のいずれに類似するかを解析するための相同性解析システムであって、前記解析対象データ群と第１のデータ群のそれぞれのデータ群に含まれるデータの相同性を示す第１の相同値ｘを算出するものであって、前記しきい値Ｅをｎ個設定して各しきい値E_i（ｉ＝１，２，…ｎ）毎に第１の相同値x_iを算出する第１の相同値算出手段と、前記解析対象データ群と第２のデータ群のそれぞれのデータ群に含まれるデータの相同性を示す第２の相同値ｙを算出するものであって、前記しきい値Ｅをｎ個設定して各しきい値E_i（ｉ＝１，２，…ｎ）毎に第２の相同値y_iを算出する第

２の相同値算出手段と、第１の相同値x_i及び第２の相同値y_i並びにしきい値の数ｎとの関係に基づいて、前記解析対象データ群が第１のデータ群又は第２のデータ群のいずれに類似するかを判定する相同性決定手段とを具備してなることを特徴とする相同性解析システムが提供される。

【０００６】

このような構成によれば、多数のデータからなるデータ群同士の相同性、すなわち類似性をまとめて評価できる。また、データとして遺伝子データを適用することにより、多数の遺伝子群の相同性をまとめて評価できる。よって、従来のように系統樹を描くために代表する遺伝子を選択する必要が無くなるため、遺伝子群としての特徴をより正確に理解することが可能となる。

【０００７】

例えば、従来は多数の遺伝子群間の関係を調べる方法では、その遺伝子群を代表する遺伝子を数個選び出し、その系統樹を描く方法であった。しかし、この方法では選び出した遺伝子が本当に代表として適当であるか判別することができず、選び出した遺伝子が適切か否かで結果が大きく異なっていた。これに対して、対象とするデータ群であるすべての遺伝子（ＯＲＦ）について相同性検索を行い、あるしきい値を達している遺伝子（ＯＲＦ）数を数えるＢＬＡＳＴ法やＦＡＳＴＡ法によれば、互いの関係を推測することができる。しかしながら、この方法でも、遺伝子群にいくつかの異なった由来の遺伝子群が混じっていた場合には、しきい値の基準を変えると結果がかなり異なってしまう特徴を有していた。そこで、本発明のように、いくつものしきい値で相同性検索を行い、それぞれのしきい値において相同性のある遺伝子（ＯＲＦ）の数をプロットし、遺伝子群間の関係を推測する方法を取ることにより、しきい値の適切性を考慮することなく極めて安定した相同性評価結果を得ることができる。

【０００８】

現在、ヒトゲノム解明以後の大きな流れとして、病気の原因となる遺伝子の選別特定が重要となってきている。候補となった遺伝子を実際のヒトＤＮＡ試料からＰＣＲ遺伝子増幅法により、ＤＮＡ断片として作成し、大腸菌などの発現系でその遺伝子が合成するタンパク質を実際に作らせ、その活性をみることや、ある

いはそのタンパク質に対する抗体を作成することが行われる。抗体は実際にヒトの病理切片などで試すことで病気の検出方法などに有用である。また、病気の原因となる遺伝子の特定は患者の組織の病域の特定、あるいは健常な人の遺伝子を解析することで、その病気にかかりやすい人を選別することができる。そうすることで、患者に対する治療の方法、例えば、保因者とそうでない人に対しては、例えばポリープ切除の緊急性の判断に応用できる。健常人に関しては、その病気に関する診断検査のそれぞれの人の遺伝子情報に基づき使い分けることが可能である。

【０００９】

また、システム（装置）に係る本発明は、その装置により実現される方法の発明としても成立する。

【００１０】

また、装置または方法に係る本発明は、コンピュータに当該発明に相当する手順を実行させるための（あるいはコンピュータを当該発明に相当する手段として機能させるための、あるいはコンピュータに当該発明に相当する機能を実現させるための）プログラム、このプログラムを記録したコンピュータ読取り可能な記録媒体としても成立する。

【００１１】

【発明の実施の形態】

本発明の骨子は、解析対象となるデータ群（解析対象データ群）と、他のデータ群との相同性を算出し、解析対象データ群がいずれのデータ群と相同性を有するか否かを判定する点にある。

【００１２】

以下、図面を参照しながら本発明の一実施形態を説明する。

【００１３】

図１は本発明の第１実施形態に係る遺伝子解析システム１０の全体構成を示す図である。図１に示すように、プロセッサ１と、このプロセッサ１に接続された出力手段２と、プロセッサ１に接続されたアミノ酸配列記憶手段３と、相同値記憶手段４から構成される。プロセッサ１は、ＢＬＡＳＴ解析手段１１とＴ検定手

段１２を備える。

【００１４】

プロセッサ１が図示しない記録媒体から所定のプログラムを読み出すことにより各手段１１及び１２としてプロセッサ１が機能する。ＢＬＡＳＴ解析手段１１は、複数のデータからなるあるデータ群と他のデータ群の相同性を相同値として算出する相同値算出手段を備える。また、Ｔ検定手段１２は、相同性決定値算出手段、相同性可否決定値算出手段及び判定結果導出手段を備える。相同性決定値算出手段は、相同性、すなわちあるデータ群と他のデータ群の類似性を数値により示した相同性決定値Z_i（１）を算出する。相同性可否決定値算出手段は、相同性決定値算出手段で算出された相同性決定値Z_i（１）が相同性を示す値として適切であるか否かを判定するための数値である相同性可否決定値Ｚ（２）を算出する。判定結果導出手段は、相同性決定値算出手段で算出された相同性決定値Z_i（１）と相同性可否決定値算出手段で算出されたに相同性可否決定値Ｚ（２）に基づいていずれのデータ群に相同性の多いデータを有するかの判定を行う。

【００１５】

次に、図２を用いて本実施形態に係る遺伝子解析に必要な遺伝子データを取得するプロセスを説明する。

【００１６】

まず、インターネットなどのネットワークを介して、遺伝子解析システム１０のＢＬＡＳＴ解析手段１１が酵母ＯＲＦ（Open Reading Frames）アミノ酸配列を取得する（ｓ２１）。次に、この取得したアミノ酸配列を機能別４３のカテゴリーに分類する（ｓ２２）。そして、得られた機能別のカテゴリーから、ミトコンドリアからの遺伝子移行の影響を受けていない酵母ＯＲＦを抽出する（ｓ２３）。なお、原核生物として、細胞寄生性の原核生物を除く原核生物のＯＲＦを１５種類取得する。

【００１７】

以上により、真核生物である酵母のＯＲＦと、原核生物の１５種類のＯＲＦが得られる。得られた各ＯＲＦは、アミノ酸配列記憶手段３に格納される。なお、これら各ＯＲＦ（遺伝子データ）は、例えば酵母であれば酵母全体の各ＯＲＦが

ひとまとまりとして遺伝子データ群として格納され、例えば原核生物であれば、ある原核生物全体の各ＯＲＦ（遺伝子データ）がひとまとまりとして遺伝子データ群として格納される。また、各ＯＲＦはアミノ酸配列により特定される。このアミノ酸配列記憶手段３に格納された遺伝子データ群の概念図を図３に示す。図３では、酵母ＯＲＦは全部でｐ個、古細菌ＡのＯＲＦはn_1個、古細菌ＢのＯＲＦはn_2個、真正細菌ＣのＯＲＦはn_3個、真正細菌ＤのＯＲＦはn_4個示してある。

【００１８】

次に、ＢＬＡＳＴ解析手段１１は、以上のようにして得られた１６種類のＯＲＦのうち、解析対象として酵母のＯＲＦを選択し、この解析対象である酵母ＯＲＦと１５種類のＯＲＦのそれぞれとの相同値ヒット数（オルソロガス遺伝子数）をＢＬＡＳＴ法を用いて算出する。

【００１９】

オルソロガス遺伝子数を検出するには、まず酵母の全ＯＲＦと各細菌の全ＯＲＦに対してＯＲＦ間の相同性の指標であるＥ-値（相同性と負の相関）を算出する。酵母と各細菌でお互いに最小のＥ-値を持つＯＲＦをオルソロガス遺伝子と判定する。オルソロガス遺伝子の数（ヒット数）が生物間の相同性を表す。そして、様々なしきい値（Ｅ-値）に対してこのヒット数を算出する。

【００２０】

図４はＢＬＡＳＴ法の概念図である。図４に示すように、酵母のＯＲＦを細胞寄生性の原核生物を除く１５種類の原核生物のＯＲＦに対して相同性検索を行う。相同性検索は、まずプロセッサ１がアミノ酸配列記憶手段３から酵母のＯＲＦ１と古細菌A_1のＯＲＦを読み出す。そして、相同性の指標であるＥ-値を算出し、最小のＥ値を持つＯＲＦを古細菌A_1の酵母ＯＲＦ１との第一ヒットＯＲＦとし、そのＯＲＦ名とその時のＥ値を記憶させる。これを全ＯＲＦに対して繰り返す。他の古細菌A_2，…，A_n、真正細菌に対しても同様に行う。ＢＬＡＳＴ法については、例えば２００１年２月１日発行の“Nature Cell Biology, Vol.3, No.2, pp210-214”に開示されている。また、遺伝子を比較する方法はＢＬＡＳＴ法以外でも、アミノ酸配列同士、塩基配列同士、分子レベルなど、遺伝子を情報

として特定できるいかなるレベルで比較してもよい。

【００２１】

さらに、図５はオルソログス遺伝子数で表される生物間の相同性の判定の概念を示す図である。図５では、矢印の先は第一ヒット遺伝子である。但し、互いのＥ値はしきい値を超えている。丸印で示したのは、互いにオルソログス遺伝子であることを示している。一方、酵母ＯＲＦ２にとっての第一ヒット遺伝子は古細菌A_1のＯＲＦ３であり、古細菌A_1のＯＲＦ３にとっての第一ヒット遺伝子は酵母ＯＲＦ３である。従って、酵母ＯＲＦ２の第一ヒット遺伝子と古細菌A_1のＯＲＦ３の第一ヒット遺伝子は一致しない。オルソログス遺伝子の数が酵母と古細菌A_1との生物間の相同性を表す。

【００２２】

図５のように、酵母と古細菌A_1に着目すると、この酵母のＯＲＦ１の古細菌A_1との第一ヒットＯＲＦ-aとそのときのＥ-値を読み出し、そして古細菌A_1のＯＲＦ-aの酵母との第一ヒットＯＲＦ-αとそのときのＥ-値を読み出す。ＯＲＦ１とＯＲＦ-αが一致し、両Ｅ-値がしきい値Ｅより大きいとき、オルソログス遺伝子と判定し、生物間相同値ヒット数としてカウントする。次に、酵母のＯＲＦ２の古細菌A_1との第一ヒットＯＲＦ-bとそのときのＥ-値を読み出し、同様に古細菌A_1とのオルソログス遺伝子の有無を判定する。ＯＲＦ-bがオルソログス遺伝子のとき、生物間相同値ヒット数としてカウントする。

【００２３】

このようにして、酵母ＯＲＦ３～酵母ＯＲＦpについても同様にオルソログスＯＲＦの有無を判定する。そして、得られたカウント値を古細菌A_1との相同値ヒット数ｘとして相同値記憶手段４に格納する。

【００２４】

以上は、酵母と古細菌A_1に着目して説明したが、同様に酵母と古細菌A_2，…A_nについても行い、それぞれを酵母との相同値ヒット数ｘとしてそれぞれ相同値記憶手段４に格納する。

【００２５】

また、プロセッサ１のＢＬＡＳＴ解析手段１１は、古細菌A_1，…，A_nについ

ての相同値ヒット数の平均値ｘ＿を算出し、得られた平均値を相同値記憶手段４に格納する。

【００２６】

また、このような相同値ヒット数の算出を他の古細菌遺伝子データ群や、真正細菌遺伝子データ群などについても行う。

【００２７】

また、ｐ個の遺伝子数である酵母遺伝子データ群とn_A個の古細菌の遺伝子データ群とのオルソロガス遺伝子検出において、あるしきい値E_iについて、ｊ番目の古細菌について得られた相同値ヒット数をx_{ij}とする。

【００２８】

また、ｐ個の遺伝子数である酵母遺伝子データ群とn_B個の真正細菌の遺伝子データ群とのオルソロガス遺伝子検出において、あるしきい値E_iについて、ｋ番目の真正細菌について得られた相同値ヒット数をy_{ik}とする。

【００２９】

以上のようにして得られた相同値ヒット数としきい値Ｅとの関係を示したのが図６である。横軸は－ｌｏｇＥ、縦軸は相同値ヒット数である。なお、この図６では、それぞれの折れ線グラフが酵母遺伝子と比較する各細菌に対応している。類似性の度合いを厳しく、すなわちしきい値Ｅを小さくしていくにつれ、条件を満たす遺伝子数、すなわち相同値ヒット数は漸減しているのが分かる。

【００３０】

以上により取得された各種相同値ヒット数に基づき、プロセッサ１のＴ検定手段は、以下によるＴ検定処理を行う。

【００３１】

Ｔ検定処理は、相同性決定値Z_i（１）の算出、相同性可否決定値Ｚ（２）の算出及び判定結果導出の３つの処理からなる。

【００３２】

まず、相同性決定値Z_i（１）の算出手法について酵母遺伝子データ群が古細菌遺伝子データ群と真正細菌遺伝子データ群のいずれに相同性が高い遺伝子を多く共有するかを解析する場合について説明する。

【００３３】

一例として、上述した各しきい値E_i（ｉ＝１，２，…，ｎ）において、有意水準５％の片側Ｔ検定をヒット数が少なくとも５以上という基準を満たす領域で行う。

【００３４】

そして、相同性決定値Z_i（１）を以下の等式により算出する。

【００３５】

【数４】

$$Z_i^{(1)} = \frac{\overline{x_i} - \overline{y_i}}{u_i} \cdot \sqrt{\frac{n_A \cdot n_B}{n_A + n_B}} \qquad (i = 1,2,\ldots,n)$$

【００３６】

なお、以上に示す等式で、x_i__はｉ番目のＥ値でのヒット数x_{ij}（ｊ＝１，２，…，n_A）の平均値を、y_i__はｉ番目のＥ値でのヒット数y_{ik}（ｋ＝１，２，…，n_B）の平均値を示す。ｎは、ＢＬＡＳＴ法で解析した際のしきい値の数を示している。また、不偏分散u_iは以下の等式で示される。

【００３７】

【数５】

$$u_i = \sqrt{\frac{1}{n_A + n_B - 2}\left\{\sum_{j=1}^{n_A}(x_{ij} - \overline{x_i})^2 + \sum_{k=1}^{n_B}(y_{ik} - \overline{y_i})^2\right\}}$$

【００３８】

以上のようにして得られた相同性決定値Z_i（１）を統計処理したのが図７である。図７では、横軸に－ｌｏｇＥを、縦軸に相同性決定値Z_i（１）をとってあり、図６に示されるデータに基づいて得られた値である。相同性有りと判断された遺伝子数を５以上の領域で棒グラフにより示している。相同性決定値Z_i（１）が$t_{nA+nb-2}$（０，１０）（＝１．７７１）以上又は－$t_{nA+nb-2}$（０，１０）（＝－１．７７１）以下の場合は有意水準５％（２つの母平均が等しいと仮定したとき、標本平均が等しくない確率が５％以下）で、各々のしきい値Ｅに対し

て古細菌あるいは真正細菌の遺伝子群が出芽酵母の遺伝子群と類似性が高い。

【００３９】

次に、相同性可否決定値Ｚ（２）の算出手法について説明する。この相同性可否決定値Ｚ（２）の算出は、１回目のＴ検定、すなわち相同性決定値$Ｚ_{i}$（１）の算出における相同性決定値$Ｚ_{i}$（１）が、$ｔ_{nA+nB-2}$（０，１０）（＝１．７７１）より設定した$Ｅ_{i}$全体に対して優位に大きいか、$－ｔ_{nA+nB-2}$（０，１０）（＝－１．７７１）より小さいか、あるいはどちらの傾向も無いかを判定するために行われる。そのため、自由度ｎ－１において以下の等式により、相同性可否決定値Ｚ（２）と$ｔ_{n-1}$（０，１０）を計算する。

【００４０】

【数６】

$$Z^{(2)} = \frac{|\overline{Z^{(1)}}| - t_{nA+nB-2}(0.10)}{s/\sqrt{n-1}}$$

【００４１】

ここで、ｓは相同性決定値$Ｚ_{i}$（１）の標準偏差、$Ｚ_{i}$（１）＿は$Ｚ_{i}$（１）の平均値である。相同性可否決定値Ｚ（２）が基準値$ｔ_{n-1}$より大きいときは$Ｚ_{i}$（１）が$ｔ_{nA+nB-2}$よりも有意水準５％で大きいと判定できる。

【００４２】

次に、以上により算出された相同性決定値$Ｚ_{i}$（１）及び相同性可否決定値Ｚ（２）に基づいて、判定結果導出を行う。判定結果導出は、図８に示す判定テーブルを用いて行う。図８に示すように、Ｚ（２）／$ｔ_{n-1}$（０，１０）が１以上で、$Ｚ_{i}$（１）の平均値が正であれば、そのカテゴリーにおける酵母ＯＲＦ群は真正細菌よりも古細菌との方に多くの相同性を有するＯＲＦを含むと判定できる。また、Ｚ（２）／$ｔ_{n-1}$（０，１０）が１以上で、$Ｚ_{i}$（１）の平均値が負であれば、そのカテゴリーにおける酵母のＯＲＦ群は古細菌よりも真正細菌との方に多くの相同性を有するＯＲＦを含むと判定できる。また、Ｚ（２）／$ｔ_{n-1}$（０，１０）が１より小さい場合はいずれの細菌に相同性を有するかを判定できないと判定する。

【００４３】

このような構成によれば、多数の遺伝子群の相同性をまとめて評価できる。よって、従来のように系統樹を描くために代表する遺伝子を選択する必要が無くなるため、遺伝子群としての特徴をより正確に理解することが可能となる。

【００４４】

例えば、この方法を用いて真核生物の起源を探ることができる。従来は、主にｒＲＮＡと小数のタンパク質について系統樹を作成し、推測していた。そして、その推測によれば、ＤＮＡ複製、転写、翻訳等の遺伝子群は古細菌起源で、エネルギー代謝の遺伝子群はミトコンドリアの共生が起源で、その他の遺伝子群は古細菌と真正細菌のモザイク構造になっていると予想されていた。また、その他説が分かれていた。

【００４５】

そこで、酵母の持つそれぞれの機能の由来を考察するため、細胞寄生性細菌を除く１５種のバクテリアのＯＲＦ（Open Reading Frames）を、機能別に分類された酵母のＯＲＦと比較する。すなわち、上述したＢＬＡＳＴ解析、Ｔ検定の解析対象を酵母遺伝子データ群、比較対照を古細菌遺伝子データ群及び真正細菌遺伝子データ群とする。その結果、４３の機能別に分類された酵母の遺伝子群のうち、２０の遺伝子群は古細菌、又は真正細菌のどちらか一方に相同性の高い遺伝子を多く含むことが分かった。一般に、進化系統の近い生物間では相同性の高い遺伝子を多く保持していることが知られていることから、酵母遺伝子の機能カテゴリー別にそれぞれの起源を推定できる。

【００４６】

具体的にこのＢＬＡＳＴ解析及びＴ検定を行うことにより、ＤＮＡ合成及び複製、転写、翻訳、減数分裂、細胞周期調節、小胞体形成、核形成などの遺伝子群は古細菌ＯＲＦと、エネルギー代謝、各種物質代謝、細胞内への物質輸送、ストレス応答、解毒、イオンホメオスタシスの遺伝子群は真正細菌ＯＲＦと、それぞれ相同性が高い遺伝子を多く共有することが分かった。

【００４７】

従って、核（遺伝情報）に関連した遺伝子は古細菌の遺伝子が、また細胞質（ホメオスタシス）に関連した遺伝子は真正細菌の遺伝子が起源であると考えられ

る。また、このことは、どちらかの遺伝子群が後から他方の生物に置き換わったことを示す。

【００４８】

これらの結果は、真核生物の核は古細菌の真正細菌への共生に由来することを示唆する。すなわち、相同性判定により遺伝子群間の関係を推測することにより、真核生物は真正細菌へ古細菌が共生することによって誕生したと考えられる可能性が高い。

【００４９】

このことを現在有力とされる各説に応じて考えると、以下のようになる。

【００５０】

まず、真正細菌由来の遺伝子はミトコンドリア共生の結果もたらされたのではないかとする説に対しては、ミトコンドリア関連遺伝子は解析のデータから削除されており、影響しないと考えられると結論づけられる。

【００５１】

また、遺伝子水平伝達の蓄積ではないかとする説に対しては、遺伝子水平伝達は、偶然の要素によって起こる。特定の機能のみ遺伝子水平伝達が起こり、その機能の遺伝子群がそっくり入れ替わるとは考えにくいと結論づけられる。

【００５２】

また、古細菌が真正細菌に進入し、共生した結果ではないかとする説に対しては、ある特定の機能に関する遺伝子群が置き換わったということは、その遺伝子群だけが置き換わりやすい特殊な環境であったことを示す。細胞内共生による核形成のプロセスはこの条件を満たし、またミトコンドリアや葉緑体の進化的形成が細胞内共生によっていることが広く認知されていることからももっともらしいといえると結論付けられる。従って、古細菌が真正細菌に進入し、共生した結果、真核生物が生じたと考えられる。

【００５３】

具体的には、真正細菌に細胞内共生した古細菌は真正細菌が合成する代謝産物を利用し続けるにつれ、それらの合成に必要な遺伝子群を失った。一方真正細菌は細胞周期などを古細菌に支配され、最終的には遺伝子複製からタンパク質合成

までのプロセスも古細菌に依存し、それ以外の遺伝子は徐々に移行したと推測できる。このことを模式的に示したのが図９である。図９に示すように、細胞周期調節、核形成、ＤＮＡ複製、ＲＮＡ転写、小胞体形成、翻訳、リボゾーム形成などに関連する核関連の遺伝子９１は、古細菌遺伝子が起源であると考えられ、細胞への物質輸送、代謝、ストレス応答、解毒、イオンホメオスタシスなどに関連する細胞質９２は、真正細菌遺伝子を起源とする遺伝子群に関連する。また、ミトコンドリア形成、ミトコンドリア輸送などに関連するミトコンドリア９３は、ミトコンドリアの共生に由来していると考えられる。

【００５４】

このように、遺伝子群をドメインレベルにまで細分化することによる相同性解析の検出能の上昇と、祖先型遺伝子をタンパク質として再現し、その活性を確認することで、遺伝子の由来に関する推測の信頼性が格段に向上する。

【００５５】

以上説明したように、本実施形態によれば、全ゲノム配列が明らかになっている酵母とバクテリアのＯＲＦデータを用い、真核生物、真正細菌、古細菌の３つの生物群における進化的な関係を明らかにすることができる。また、ヒトやその他の生物のＯＲＦデータを用いてその相同性を解析することにより、あらゆる生物同士の相同性をしきい値の設定条件に依存せずに安定して極めて簡便に判定することができる。

【００５６】

本発明は上記実施形態に限定されるものではない。ＢＬＡＳＴ法を用いて相同値ヒット数を算出したが、他の相同値算出法を適用することも可能であることはもちろんである。

【００５７】

また、例えば、非常に多数の遺伝子集団を用いて、ある目的に従って比較したり、特定の遺伝子を選別したり、あるいは選別された遺伝子を具体的に特定することに本発明を適用することができる。すなわち、対象とする遺伝子グループが複数ある場合には、それぞれのしきい値（Ｅ値）で計算することにより、非常に多数の遺伝子を有する複数のグループ間におけるそれぞれの遺伝子の、ある特徴

を有する遺伝子群との相同性類似性を比較すること、さらにそこで対象となった遺伝子を抽出、特定することを非常に簡便に行うことができる。さらに、対象とした遺伝子群の中に、そのグループの中では特異な特徴を有する遺伝子を特定することもできる。例えば、生物の進化過程においては親から子への伝達ではない、いわゆる水平伝達で導入された遺伝子の抽出にも適用可能である。

【００５８】

特定された遺伝子を解析した生物の中から取り出し、ＤＮＡ塩基配列を知ること、あるいは類似の遺伝子を検索し、類似の遺伝子の情報に基づきそれらの遺伝子の機能を推測することが可能となる。また、ある目的に従った人材の選別などにも応用できる。すなわち、非常に多数のデータを含むグループ（データ群）の中から、ある一定の条件を満たす、又はその条件の度合いを変化させつつデータを選別、特定できる。グループ間で相互に相同性を解析することにより、お互いの条件を満たすペアを選別することもできる。

【００５９】

また、現在、ヒトゲノム解明以後の大きな流れとして、病気の原因となる遺伝子の選別特定が重要となっている。候補となっている遺伝子を実際のヒトＤＮＡ試料からＰＣＲ遺伝子増幅法により、ＤＮＡ断片として作成し、大腸菌等の発現系でその遺伝子が合成するタンパク質を実際に作らせ、その活性をみることや或いはそのタンパク質に対する抗体を作成することが行われる。抗体は実際にヒトの病理切片などで試すことで病気の検出方法等に有用である。また、病気の原因となる遺伝子の特定は患者の組織の病域の特定、或いは健常な人の遺伝子を解析することでその病気にかかりやすいヒトを選別することができる。そうすることで、患者に対する治療の方法、例えば、その保因者とそうでない人に対しては、例えばポリープ切除の緊急性の判断に応用できる。健常人に関してはその病気に関する診断検査のそれぞれの人の遺伝子情報に基づき使い分けることが可能である。

【００６０】

また、相同性比較の対象とするデータ群として、遺伝子以外のデータを用いることにより、人材情報データの解析、選別や、商品の類似性判定、気象情報解析

などにも応用できる。すなわち、複数のデータからなるデータ群同士の類似性の度合いを解析するものであれば何でも適用可能である。

【００６１】

【発明の効果】

以上詳述したように本発明によれば、多数の遺伝子同士を定量的にかつ正確に比較することが可能となる。

【図面の簡単な説明】

【図１】

本発明の第１実施形態に係る遺伝子解析システムの全体構成を示す図。

【図２】

同実施形態に係る遺伝子解析に必要な遺伝子データを取得するプロセスの一例を示す図。

【図３】

同実施形態に係るアミノ酸配列記憶手段に格納された遺伝子データ群の概念図。

【図４】

同実施形態に係るＢＬＡＳＴ法の概念図。

【図５】

同実施形態に係るＢＬＡＳＴ法による相同性の有無の判定の概念を示す図。

【図６】

同実施形態に係る相同値ヒット数としきい値Ｅとの関係を示す図。

【図７】

同実施形態に係る相同性決定値Z_i（１）を示す図。

【図８】

同実施形態に係るＴ検定における判定結果導出に用いられる判定テーブルの一例を示す図。

【図９】

同実施形態に係るＴ検定に於ける判定結果により考察した酵母遺伝子の由来を模式的に示す図。

【符号の説明】

１…プロセッサ

２…出力手段

３…アミノ酸配列記憶手段

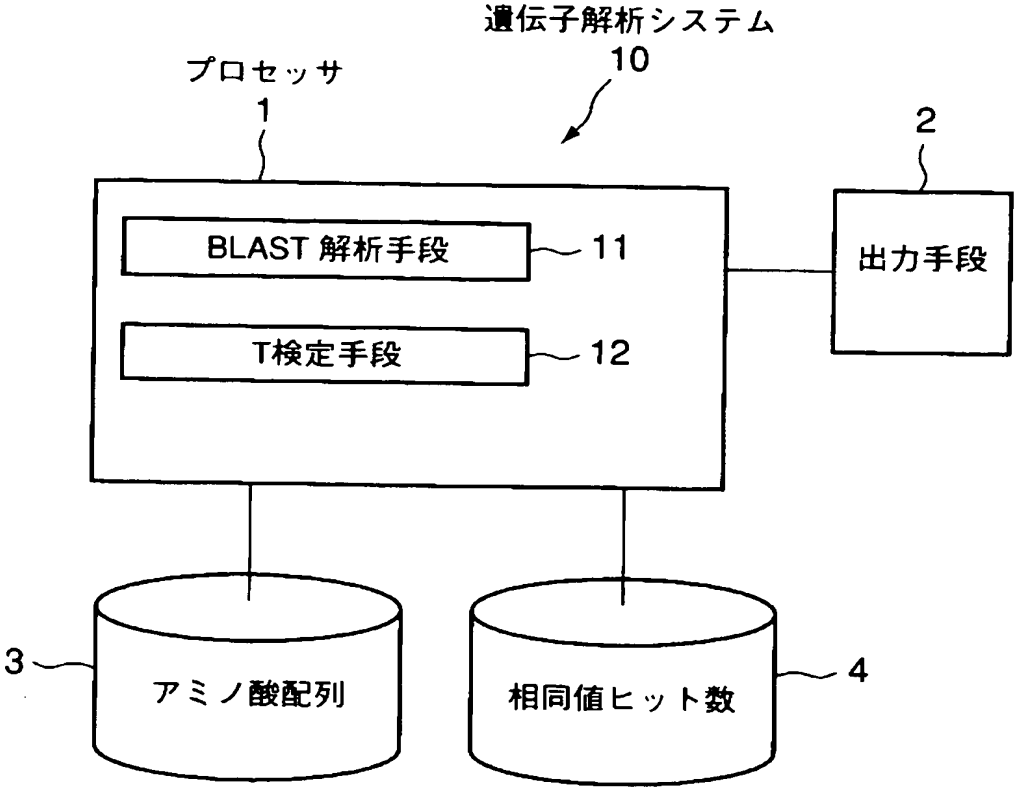
４…相同値記憶手段

１１…ＢＬＡＳＴ解析手段

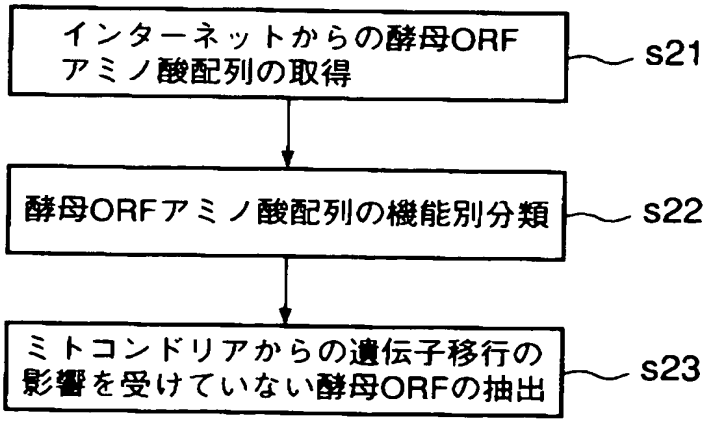
１２…Ｔ検定手段

【書類名】　図面

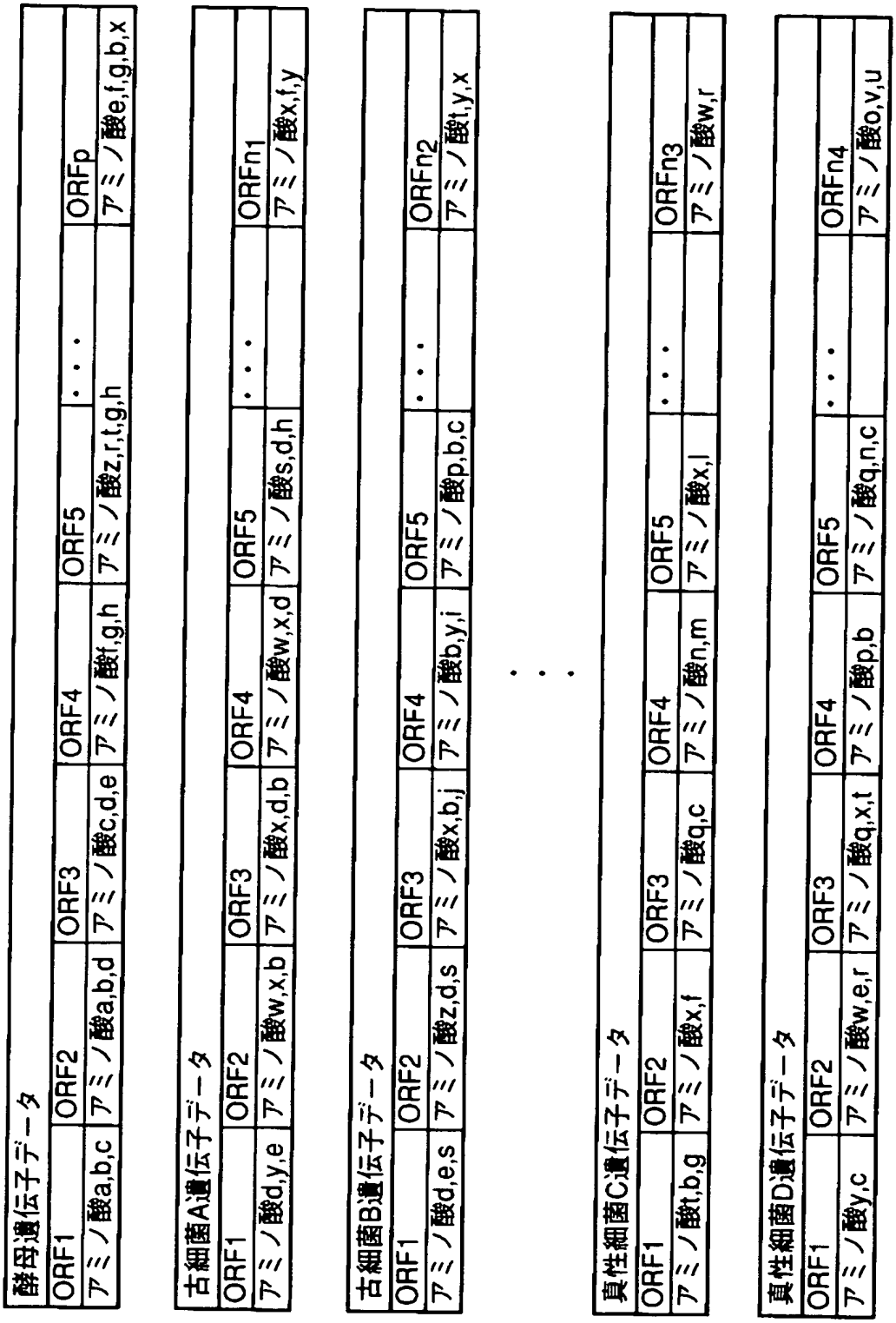
【図１】


遺伝子解析システム
10
プロセッサ
1
2
BLAST 解析手段
11
出力手段
T検定手段
12
3
アミノ酸配列
相同値ヒット数
4

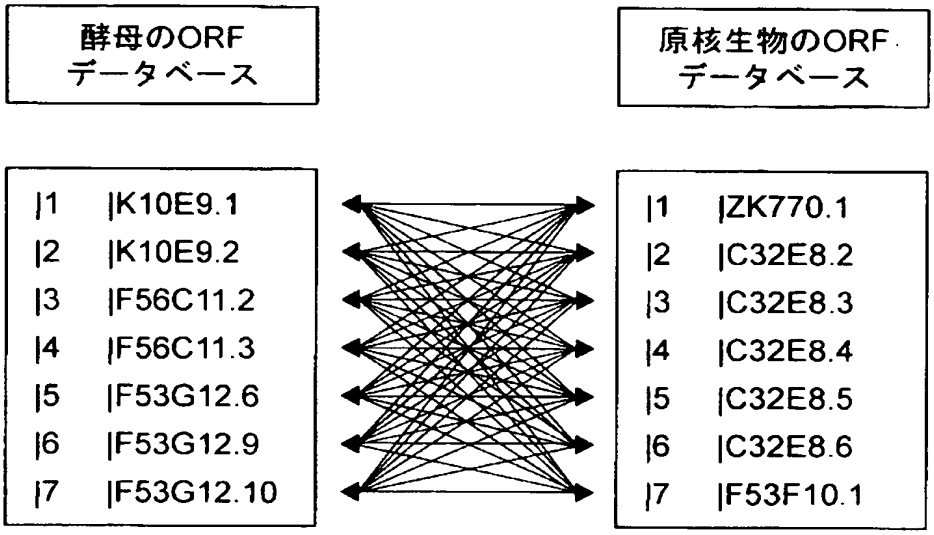
【図２】


インターネットからの酵母ORF
アミノ酸配列の取得
s21
酵母ORFアミノ酸配列の機能別分類
s22
ミトコンドリアからの遺伝子移行の
影響を受けていない酵母ORFの抽出
s23

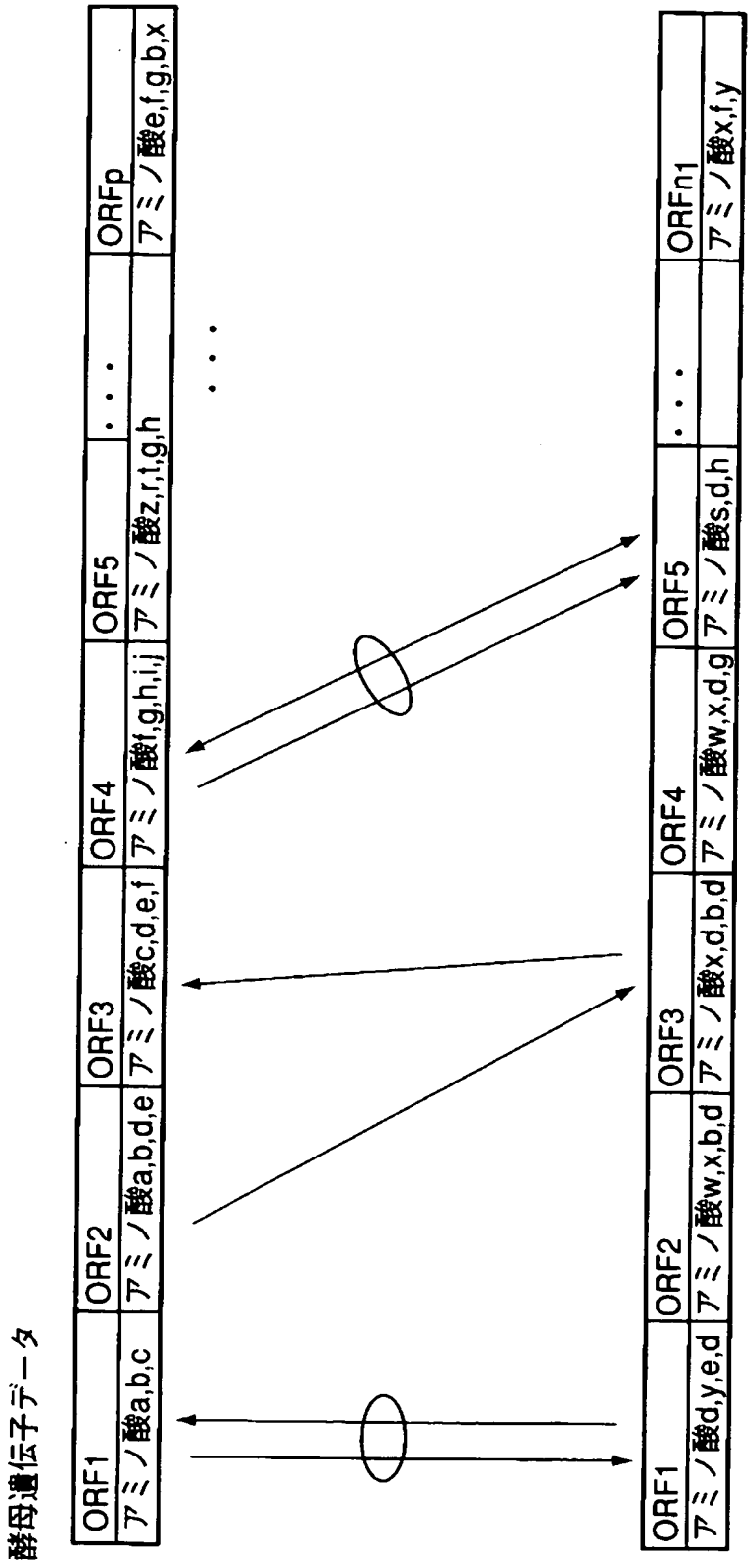
【図３】


酵母遺伝子データ
ORF1
ORF2
ORF3
ORF4
ORF5
・・・
ORFp
アミノ酸a,b,c
アミノ酸a,b,d
アミノ酸c,d,e
アミノ酸f,g,h
アミノ酸z,r,t,g,h
アミノ酸e,f,g,b,x
古細菌A遺伝子データ
ORF1
ORF2
ORF3
ORF4
ORF5
・・・
ORFn1
アミノ酸d,y,e
アミノ酸w,x,b
アミノ酸x,d,b
アミノ酸w,x,d
アミノ酸s,d,h
アミノ酸x,f,y
古細菌B遺伝子データ
ORF1
ORF2
ORF3
ORF4
ORF5
・・・
ORFn2
アミノ酸d,e,s
アミノ酸z,d,s
アミノ酸x,b,j
アミノ酸b,y,i
アミノ酸p,b,c
アミノ酸t,y,x
真性細菌C遺伝子データ
ORF1
ORF2
ORF3
ORF4
ORF5
・・・
ORFn3
アミノ酸t,b,g
アミノ酸x,f
アミノ酸q,c
アミノ酸n,m
アミノ酸x,l
アミノ酸w,r
真性細菌D遺伝子データ
ORF1
ORF2
ORF3
ORF4
ORF5
・・・
ORFn4
アミノ酸y,c
アミノ酸w,e,r
アミノ酸q,x,t
アミノ酸p,b
アミノ酸q,n,c
アミノ酸o,v,u

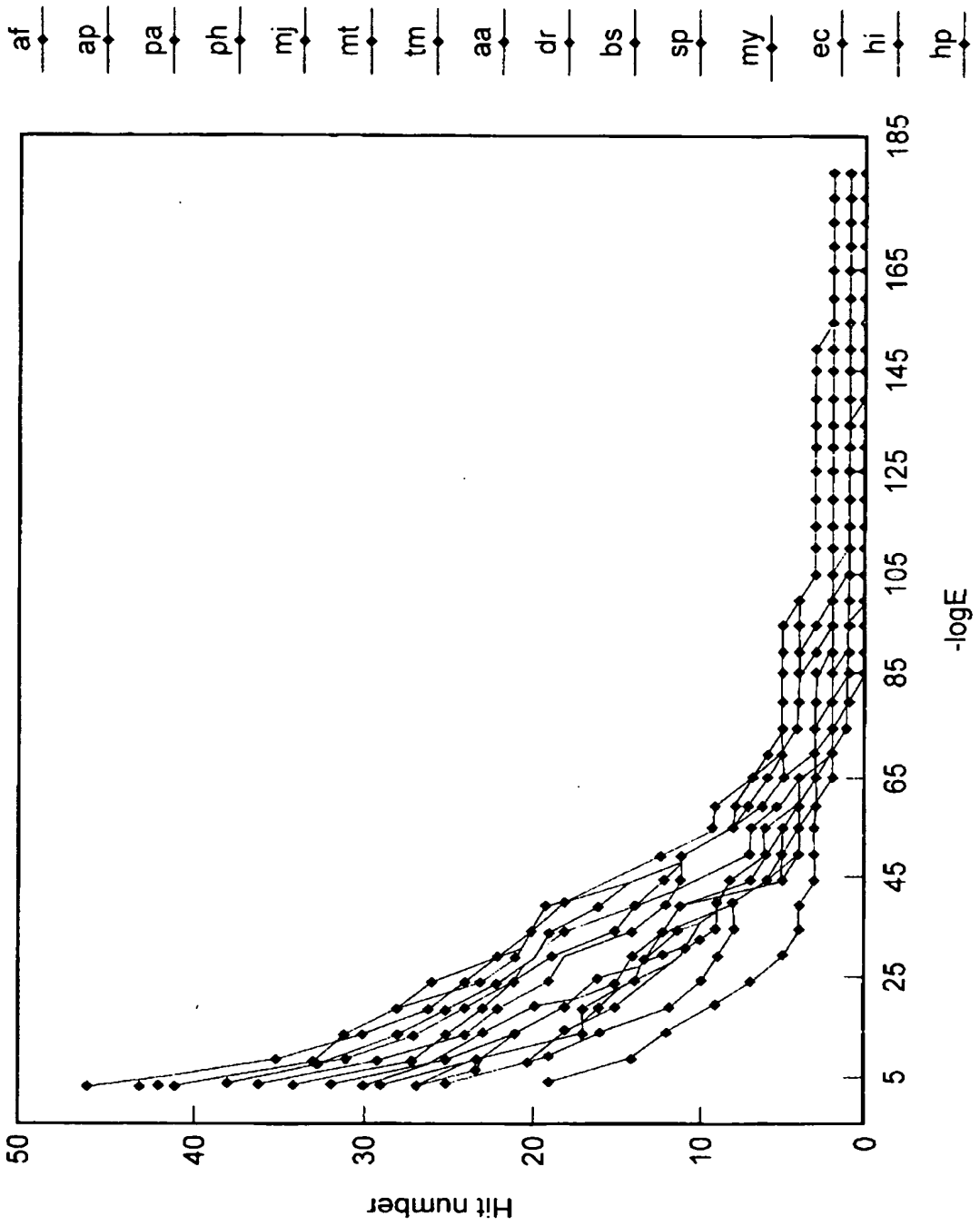
【図４】


酵母のORF
データベース
原核生物のORF
データベース
|1 |K10E9.1
|2 |K10E9.2
|3 |F56C11.2
|4 |F56C11.3
|5 |F53G12.6
|6 |F53G12.9
|7 |F53G12.10
|1 |ZK770.1
|2 |C32E8.2
|3 |C32E8.3
|4 |C32E8.4
|5 |C32E8.5
|6 |C32E8.6
|7 |F53F10.1

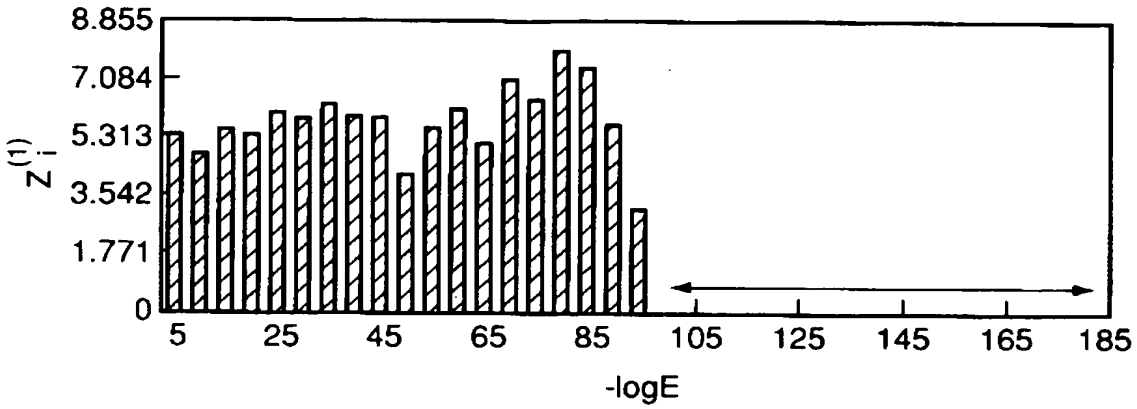
【図５】


酵母遺伝子データ
ORF1
アミノ酸a,b,c
ORF2
アミノ酸a,b,d,e
ORF3
アミノ酸c,d,e,f
ORF4
アミノ酸t,g,h,i,j
ORF5
アミノ酸z,r,t,g,h
・・・
ORFp
アミノ酸e,f,g,b,x
古細菌A1の遺伝子データ
ORF1
アミノ酸d,y,e,d
ORF2
アミノ酸w,x,b,d
ORF3
アミノ酸x,d,b,d
ORF4
アミノ酸w,x,d,g
ORF5
アミノ酸s,d,h
・・・
ORFn1
アミノ酸x,f,y

【図６】


af
ap
pa
ph
mj
mt
tm
aa
dr
bs
sp
my
ec
hi
hp
Hit number
0
10
20
30
40
50
-logE
5
25
45
65
85
105
125
145
165
185

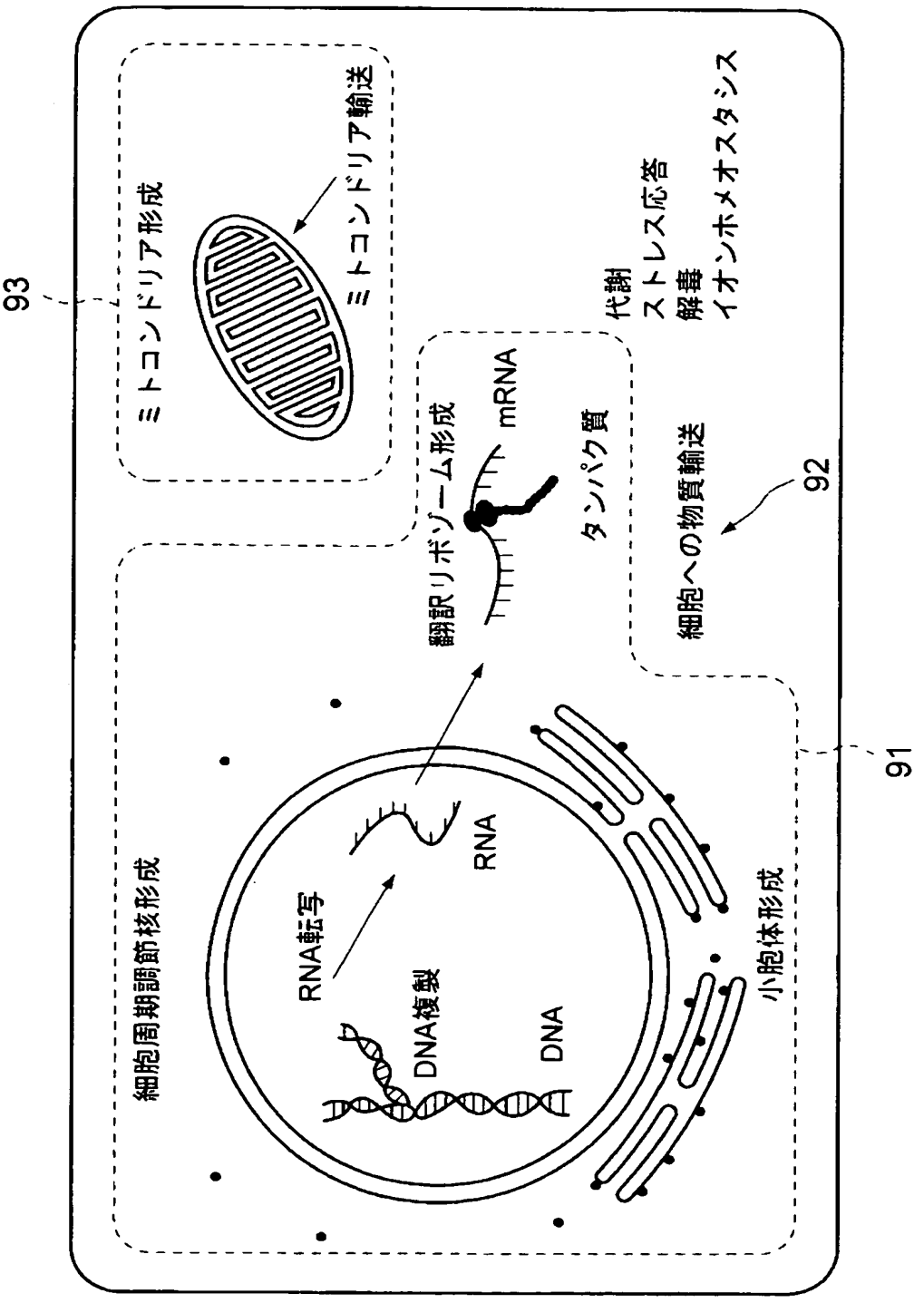
【図７】



【図８】

$Z^{(2)}/t_{n-1}$ ＼ $\overline{Z^{(1)}}$	＋	－
1以上	古細菌	真正細菌
1未満	判定不可	判定不可

【図９】


93
ミトコンドリア形成
ミトコンドリア輸送
92
代謝
ストレス応答
解毒
イオンホメオスタシス
翻訳リボゾーム形成
mRNA
タンパク質
細胞への物質輸送
91
細胞周期調節核形成
RNA転写
RNA
DNA複製
DNA
小胞体形成

【書類名】　　要約書

【要約】

【課題】多数の遺伝子同士を定量的にかつ正確に比較する。

【解決手段】ＢＬＡＳＴ解析手段１１により、解析対象データ群と第１のデータ群のそれぞれのデータ群に含まれるデータの相同性を計算し、その値を用いて相同性有りとなるデータの数（オルソロガス遺伝子数）を第１の相同値ｘとして算出し、しきい値Ｅをn_A個設定して各しきい値E_i毎に第１の相同値x_iを算出し、解析対象データ群と第２のデータ群のそれぞれのデータ群に含まれるデータの相同性を計算し、相同性有りとなるデータの数（オルソロガス遺伝子数）を第２の相同値ｙとして算出し、しきい値Ｅをn_B個設定して各しきい値E_i毎に第２の相同値y_iを算出し、Ｔ検定手段１２は、第１の相同値x_i及び第２の相同値y_i並びにしきい値の数ｎとの関係に基づいて、解析対象データ群が第１のデータ群又は第２のデータ群のいずれに類似するかを判定する。

【選択図】　　図１

特願２００１－１８３８５６

出　願　人　履　歴　情　報

識別番号　　［３９６０２０８００］

１．変更年月日　　１９９８年　２月２４日
［変更理由］　　名称変更
住　所　　埼玉県川口市本町４丁目１番８号
氏　名　　科学技術振興事業団